Stage-wise Training Strategy for Multiple Mental Disorder Detection: Focused KAN in Transformers

Team 5:
Fangjing Wu (fw179)
Giselle Qiu (xq55)
Minxuan Hong (mh1653)

I. INTRODUCTION

Mental disorders significantly impact daily life but are challenging to detect due to their subjective and often inarticulate nature. Diagnoses typically rely on patient self-reports and clinicians' expertise, utilizing standardized scales like the GAD-7. However, these methods are prone to biases and concealment, as individuals may withhold critical information or downplay symptoms during consultations. Traditional scales also struggle to capture the nuanced linguistic information present in patients' language. Machine learning models offer a promising solution by identifying subtle linguistic cues and leveraging alternative data sources such as social media posts and comments. These models can mitigate biases inherent in face-to-face interactions by analyzing patients' daily habits, which are usually inaccessible in conventional care settings (Le Glaz et al., 2021). As Le Glaz and colleagues note, "machine learning and NLP techniques provide useful information from unexplored data." Nonetheless, ethical considerations must be addressed to ensure these methods support rather than replace clinical practice.

Real-life data is crucial for enabling models to learn meaningful patterns. Research indicates that social media posts—spontaneous expressions of personal feelings and symptoms—are valuable for text-based mental disorder detection (Montejo-Ráez et al., 2024). However, most studies focus on single disorders like anxiety or depression (Losada et al., 2017; Lee et al., 2021), despite the high prevalence of comorbidities among mental disorders (Roca et al., 2009). Curated datasets for analysis are often binary-labeled as "case" (1) or "control" (0), which overlooks distinct yet overlapping features necessary for detecting nuanced or comorbid conditions. According to the NIH, 23.1% of U.S. adults experienced Any Mental Illness (AMI) in 2022, with 3% having multiple concurrent mental illnesses, highlighting the importance of addressing comorbidities in research and clinical practice.

Neural networks are well-suited for this task, excelling at capturing interactive features without explicit interaction terms. Modern large-scale models using attention mechanisms have shown remarkable performance in natural language processing (Vaswani et al., 2017; Brown et al., 2020; Devlin et al., 2019). By focusing on salient information flows, these models hold great potential for improving mental disorder detection. However, training multi-task models to classify multiple mental disorders simultaneously presents challenges. High-quality, multi-labeled datasets are scarce, as most existing datasets are binary-annotated for single disorders and vary significantly in size. Combining these datasets can lead to severe class imbalances, causing models to favor larger datasets while neglecting smaller ones. Additionally, each mental disorder has unique linguistic and contextual characteristics, and mixed datasets might result in overly generic feature extraction, reducing task-specific accuracy. Varying task difficulties can also introduce biases, with models overfitting simpler tasks and underfitting more complex ones.

To address these challenges, we propose a novel stage-wise training strategy based on a modified neural network architecture. This approach optimizes learning across tasks while accounting for data imbalances, unique disorder characteristics, and varying task complexities.

II. METHODOLOGY

The initial model consists of one embedding layer, Transformer modules with one multi-head attention layer each, feed-forward KAN layers, and a binary output layer.

2.1 Transformer Architecture with Kolmogorov-Arnold Network

In our study, we enhanced the Transformer architecture by replacing the traditional multi-layer perceptron (MLP) in the feed-forward layers with a Kolmogorov-Arnold Network (KAN) (Liu et al., 2024). By parameterizing each edge weight as a learnable, nonlinear spline function, KAN effectively captures complex functional structures and improves approximation capabilities. This modification offers significant advantages in accuracy and interpretability, particularly excelling in small-scale, information-rich tasks. Additionally, KAN reduces model size and computational power requirements, making it ideal for training large models on limited datasets, such as those used for mental disorder detection. To further enhance our model's ability to capture intricate linguistic patterns, we incorporated multi-head attention layers into the KAN-based architecture. These attention mechanisms are crucial for focusing on significant tokens within the text, with each attention head specializing in distinct features. This specialization enables the model to identify subtle cues relevant to specific disorders while maintaining a comprehensive understanding of both broad contexts and fine details. The integration of KAN with multi-head attention thus provides a robust, scalable solution for detecting nuanced and comorbid mental health conditions, balancing interpretability with high performance.

2.2 Regularization and Dropout

To enhance generalization and prevent overfitting, our model integrates regularization and dropout techniques. Within the KAN layers, a custom loss penalizes the L1 norm of spline weights and includes an entropy-based penalty, promoting smooth weight updates and balanced activation across spline functions. Additionally, dropout is extensively applied throughout the architecture: after the softmax operation in multi-head attention mechanisms and after both self-attention and feed-forward layers in Transformer modules, with a default rate of 0.3. This introduces stochasticity by randomly deactivating neurons during training, further mitigating over-reliance on specific features and enhancing the model's robustness.

2.3 Stage-Wise Training Strategy with Attention Head Freezing

To enhance model performance for multiple mental disorder detection, we introduce a stage-wise training strategy that processes each disorder-specific dataset sequentially using a greedy algorithm for independent binary classification tasks. A key component of this strategy is attention head freezing, which mitigates catastrophic forgetting by preserving task-specific knowledge. During each training stage, we monitor weight changes across all attention heads; once changes fall below a predefined threshold, the corresponding head is frozen to retain learned features. This allows the model to incorporate new tasks without disrupting previously acquired knowledge.

After achieving satisfactory performance on a task, the model advances to the next dataset. We track the number of active attention heads (total heads minus frozen ones) and introduce additional heads if the active count falls below a minimum threshold, ensuring continued flexibility for learning new information. Upon training on all individual tasks, the model undergoes a final fine-tuning phase with a multi-label dataset. In this phase, the loss function switches to binary cross-entropy, and the output layer is adjusted for five labels. All attention heads are frozen, and only the embedding and feed-forward layers are updated, consolidating specialized representations for improved multi-disorder classification.

This multi-stage strategy with dynamic attention head freezing ensures modularity, preserves knowledge across tasks, and effectively prevents catastrophic forgetting. By focusing on one task at a time and selectively freezing attention heads, the model maintains task-specific precision while enabling seamless knowledge transfer and integration of shared features. Additionally, incorporating a composite data replay mechanism with a subset of previous task data further reduces forgetting, enhancing the model's flexibility and scalability for continuous learning and fine-tuning as new data becomes available.

2.4 Multi-Label Classification Finetune

During the finetuning phase, the binary pre-trained model was adapted to the new multi-label classification task by adding an additional output layer and making subtle adjustments to all weights. This output layer was specifically designed to convert binary predictions into multi-label outputs with regard to the presence of co-occurring mental disorder conditions. To ensure smooth and stable adaptation to the multi-label task, training was conducted with a lower learning rate, preventing drastic changes to the pre-trained weights. To address the issue of class imbalance in the dataset, a dynamic thresholding function was implemented, optimizing thresholds for each label based on the best F1 score achieved during training. This dynamic adjustment helps improve the precision and recall of each label, with particular focus on rarer conditions such as OCD and bipolar disorder.

2.5 Model Evaluation Metrics

We evaluate the model's performance at each stage using categorical cross-entropy loss and F1 scores. The model is optimized by minimizing the discrepancy between predicted probabilities and true labels for each binary task. Once optimal performance is achieved for each disorder, we evaluate the final model using binary cross-entropy with logits loss, macro and micro F1 scores, along with the exact match ratio (EM). The model used Binary Cross-Entropy with Logits as the loss function, which is well-suited for multi-label classification tasks where multiple labels can be activated simultaneously. The EM metric provides a stricter evaluation by measuring the percentage of samples where the predicted labels exactly match the true labels.

III. Dataset Acquisition and Composition

This project relies on two main datasets: five training datasets and a synthesized multi-label dataset for fine-tuning.

3.1 Multi-stage training datasets

The training datasets were created from publicly available Reddit data, sourced from Pushshift dumps spanning 2005–2023. These datasets focus on five specific mental health conditions—depression, anxiety, bipolar disorder, OCD, and PTSD—by extracting case data from their respective subreddits (*r/depression, r/Anxiety, r/bipolar, r/OCD, and r/ptsd*). For each mental health condition, the extraction process prioritized posts from the most recent period, starting from December 31, 2023, and extending back to earlier years depending on the availability and quality of posts. Using the PTSD data as an example:

- **PTSD Dataset (r/ptsd):** PTSD-related posts were extracted from 2015 to 2023, yielding a total of 71,963 posts prior to preprocessing. To ensure the full context was captured, the data included both the titles and bodies of the posts, which were concatenated into a single text field, as each component often contains valuable information. During preprocessing, posts marked as null or empty (e.g., posts labeled as "[removed]" or "[deleted]" or those containing only URLs) were removed. After this cleaning process, 50,422 posts remained. From this cleaned dataset, a subset of **30,000** usable posts was randomly selected for analysis. Given the relatively lower volume of PTSD-related posts compared to other mental health conditions, the extraction period for PTSD was extended further back in time, from 2015 to 2023, to ensure a sufficient sample size for analysis.

To create control data, posts were randomly sampled from six general-interest subreddits (*r/ChangeMyView, r/NoStupidQuestions, r/Showerthoughts, r/books, r/movies, r/CasualConversation*). Similar to the above, for each control, the extraction process starts from December 31, 2023, and extends back to earlier years, depending on the quality and quantity of the available data. Using the control extracted from r/ChangeMyView as an example:

- **Control Dataset (r/ChangeMyView)**: The control data, consisting of titles and bodies of posts, was extracted from the subreddit r/ChangeMyView for the period from 2019 to 2023. This resulted in a total of 154,559 posts before preprocessing. Posts in this subreddit typically contain content primarily in the title, with the body often left empty. During preprocessing, null or empty titles were removed, including posts marked as "[removed]" or "[deleted]" and those containing only URLs. After cleaning, 36,723 posts remained, from which **25,000** usable cases were randomly selected for analysis.

The decision to select 25,000 posts for each control was based on maintaining the balance between control and case data. Since control data was sourced from six general-interest subreddits, and the dataset for each mental health condition contains 30,000 posts, the total control data needed to match the total case data. To achieve this balance, each mental health dataset was paired with 30,000 control posts, evenly distributed across the six subreddits (5,000 posts from each subreddit, totaling 30,000 control posts per condition). This ensures that both control and case data are equal in size for every mental health condition, enabling balanced and unbiased analysis. The datasets were split into training and testing sets with an 80-20 ratio to facilitate model training and evaluation.

3.2 Finetune multi-labeled dataset

The multi-label dataset for this project was synthesized using a Kaggle dataset comprising mental health-related statements tagged with various mental health statuses. This dataset integrates data from multiple sources and platforms, such as Reddit and Twitter, ensuring diversity and depth. The original dataset includes statements tagged with one of the following seven mental health statuses: “Normal”, “Depression”, “Suicidal”, “Anxiety”, “Stress”, “Bi-Polar”, and “Personality Disorder”. We further annotated the data using an automated process with ChatGPT. Texts were categorized based on their alignment with our five mental health conditions—depression, anxiety, OCD, bipolar disorder, and PTSD—following a set of predefined rules: Posts labeled as "Normal" were assigned zeros for all conditions, while those labeled with mental health issues were further annotated based on keywords or logical mappings. In cases where no explicit condition matched, the label "depression" was assigned as a default. This dataset contains 53,044 posts, with 7.42% annotated with two or more condition labels, making it suitable for multi-label classification tasks.

IV. RESULTS

4.1 Stage-Wise Binary Training Results

The initial model was trained sequentially on the five independent datasets. Each dataset was chosen randomly and fed into the model. During the first training process based on the depression dataset, the weights of attention heads were observed to converge (Fig 3). The convergence is evident from the decreasing oscillations in the weight changes for all four attention heads over the training steps. A threshold of a mean absolute weight change below 1e-6 was used to determine the stability of a head, ensuring that only those heads with minimal changes in weight were frozen.

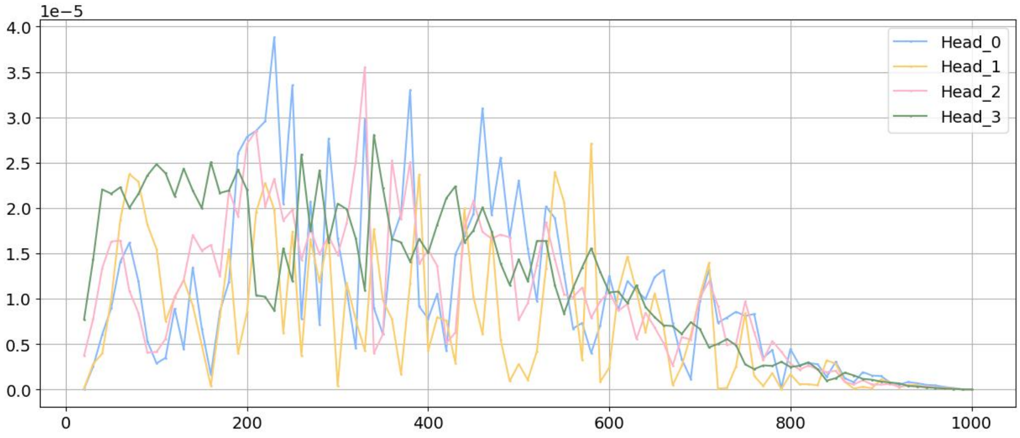

Figure 3. Attention weight changes in stage one — depression dataset training

The training results across each stage are shown in Table 3 and Figure 4. The stabilization of train loss and the validation loss means our model generalizes well without overfitting. The drop in loss from the first stage to the following stage reflects the model's transition from pre-trained BERT-mini embeddings to a more specialized state for the specific disorder.

In subsequent stages, the freezing mechanism and the introduction of new attention heads for each disorder allow the model to keep learning unique features while retaining previously learned knowledge. However, the increased loss difference in Stage 3 is likely due to the significant linguistic overlap with depression, as well as the broader, more nuanced expressions of anxiety. Since anxiety often co-occurs with other disorders, especially depression, isolating features unique to anxiety is more challenging. Overall, the similar loss values across later stages suggest comparable task complexity and that the model has learned generalizable linguistic cues. The carefully curated and balanced datasets for each disorder also ensured a uniform task difficulty.

Table 3. Stage-wise training results for each dataset

Datasets	Train Loss	Valid Loss	Recall	Precision	F1 Acc
Depression	0.208	0.241	0.957	0.965	0.961
Bipolar	0.052	0.150	0.959	0.96	0.959
Anxiety	0.052	0.159	0.959	0.96	0.968
PTSD	0.052	0.112	0.973	0.964	0.972
OCD	0.05	0.101	0.973	0.97	0.963

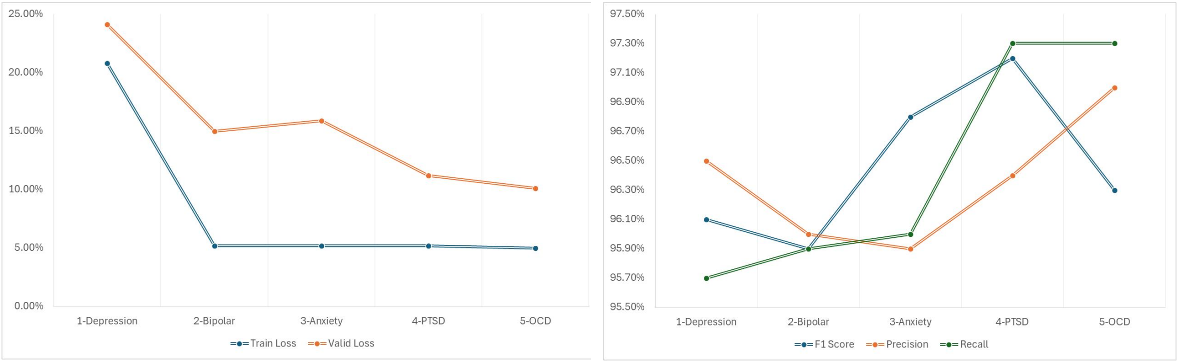

Figure 4. Stage-wise training results visualization

The consistently high precision, recall, and F1 scores indicate the model effectively captured disorder-specific features from different datasets during stage-wise training. The model tends to perform better on disorders like PTSD and OCD that have distinct and easily identifiable linguistic markers like "trauma" or "rituals". In contrast, disorders like depression and anxiety, which share overlapping patterns like "worry," "sad," or "overwhelmed", showed a slight decrease in precision and recall due to confusion between them.

4.2 Final Finetune Multi-label Results

To expand the model from binary to a multi-label task, we leverage transfer learning to finetune our final model. The model was trained on the synthesized multi-label dataset to predict the presence of multiple disorders simultaneously. The evaluation metrics include exact match ratio, which considers a multi-label prediction correct only if all assigned labels exactly match the true labels, macro F1, and micro F1 scores, which account for partially correct outputs. Due to conditions like depression and anxiety being much more common in real-world data, detecting rarer disorders like OCD and bipolar is naturally more challenging. Yet, our model was able to achieve good performance across all tasks.

Table 4. Final Finetune Multi-label Results

BCE loss		Label-wise F1 scores					Overall Model Evaluation		
training loss	validation loss	Depression	Bipolar	PTSD	OCD	Anxiety	Macro-F1	Micro-F1	EM
0.05	0.06	0.953	0.955	0.909	0.901	0.944	0.932	0.959	0.936

4.3 User-level Predicting

We further evaluated our final model using self-generated sample sentences that simulate various combinations of co-occurring disorders. The results, as shown in Table B4, underscore the model's robustness in identifying complex mental health conditions. A notable aspect of this evaluation is that the test sentences deliberately excluded explicit disorder-related keywords—such as "depressed," "anxious," or direct mentions of disorder names—that are commonly present in the training data. This approach highlights the model's ability to infer underlying conditions through nuanced linguistic and contextual cues, rather than relying solely on explicit terminology.

V. DISCUSSION

Our proposed stage-wise training approach, incorporating Kolmogorov-Arnold Networks (KANs) into Transformer-based models, has demonstrated significant effectiveness. The initial model, consisting of two Transformer modules with dimensions of 16 and 4 heads each, achieved a high accuracy of 95% in the first training stage and maintained similar performance across subsequent stages after 1,000 training steps. This strategy effectively prevented catastrophic forgetting through the use of a freezing-head mechanism. Specifically, training began with the depression dataset, followed by continual training on the bipolar dataset with and without attention head freezing. The model employing frozen attention heads exhibited slightly higher accuracy, as anticipated when training on diverse, independent datasets.

When finetuning a model previously trained solely on the depression dataset, we observed a decline in depression detection accuracy from 96% to 74%, with near-zero accuracies for bipolar disorder, PTSD, and OCD, and anxiety detection at approximately 35%. These results align with psychological studies indicating strong correlations between depression and anxiety, suggesting that similar linguistic patterns may cause model confusion. In contrast, our approach showed substantial effectiveness in predicting multiple mental disorders post-finetuning. The model with frozen attention heads and updated KANs achieved high accuracy and interpretability in multi-label classification, effectively managing comorbidities and overlapping features among disorders. This was evident as the model maintained robust performance in binary classifications during intermediate stages and excelled in final multi-label tasks.

Unlike traditional multi-task learning paradigms, our training process is structured into sequential stages, progressing from binary to multi-label classification tasks. This incremental training develops robust, disorder-specific feature representations while allowing the model to learn shared features through task-specific head freezing. As highlighted by Ruder (2017), multi-task learning often suffers from task interference due to competing representations. Our stage-wise approach mitigates this by focusing on one task at a time, preserving task-specific learning before advancing to more complex multi-label tasks. This structured progression balances shared and task-specific feature development, ensuring optimal performance across all stages.

The head-freezing strategy is pivotal in mitigating catastrophic forgetting and preserving task-specific knowledge. After training on each task, critical attention heads are frozen based on weight stability, allowing the model to learn new tasks without overwriting prior knowledge. This method, similar to Progressive Neural Networks (Rusu et al., 2016), maintains a fixed architecture by utilizing attention heads for parameter efficiency. In the final fine-tuning phase, previously frozen heads are selectively adjusted to align task-specific features with multi-label predictions, ensuring robust performance across all disorders. Integrating KANs enhances the model's interpretability and scalability by replacing fixed activation functions with learnable splines, enabling the capture of subtle, context-dependent linguistic patterns. KANs efficiently model intricate, high-dimensional relationships, outperforming conventional multi-layer perceptrons (MLPs) in handling complex linguistic and contextual nuances essential for mental health analysis.

Overall, our stage-wise training approach effectively integrates KANs with Transformer-based models, addressing multiple challenges in mental disorder prediction. The combination of stage-wise training, task-specific head freezing, and flexible KAN modeling results in high accuracy, stability, and interpretability in multi-label classification tasks. This method mitigates catastrophic forgetting and task interference while balancing task-specific and shared feature learning, outperforming traditional multi-task learning methods and demonstrating strong potential in managing comorbidity and feature overlap in mental disorder detection.

5.2 Limitations and Perspectives

Despite the effectiveness of our training strategy and model, several limitations must be acknowledged. Data-wise, reliance on Reddit and Kaggle datasets may limit generalizability due to demographic biases inherent in Reddit's user base and potential biases from automated annotations in Kaggle. Additionally, using posts from general-interest subreddits as control data could introduce biases, making it easier to differentiate between mental disorder datasets and controls.

Methodologically, employing binary cross-entropy with logits as the primary loss function presents challenges such as class imbalances, inter-disorder relationships, and difficulties in classifying disorders like PTSD and OCD. This can lead to the over-prediction of more prevalent disorders like depression at the expense of rarer conditions. To address these issues, future work could incorporate weighting schemes, focal loss, and hierarchical loss functions to better manage class imbalances and interdependencies between disorders. Adopting pre-trained models specialized in mental health applications could also enhance contextual understanding and feature extraction, providing a stronger foundation for downstream tasks.

Future research should aim to enhance data representativeness by incorporating more diverse datasets across multiple languages, cultural contexts, and demographic groups. Improving data annotation through manual validation or advanced semi-supervised techniques could mitigate existing biases. Methodological advancements may include exploring meta-learning or transfer learning to improve adaptability to new or underrepresented disorders. Additionally, integrating interpretability tools such as SHAP or attention-based visualizations would enhance model transparency and clinical applicability, fostering trust and usability in mental health applications. These efforts will collectively refine model performance and expand its utility in accurately detecting and understanding mental health conditions, ensuring that developed models are robust and applicable across diverse clinical scenarios.

VI. CONCLUSION

In this study, we propose a stage-wise training strategy for mental disorder detection by integrating Kolmogorov-Arnold Networks (KAN) into Transformer-based models, effectively replacing traditional feed-forward layers to capture complex linguistic patterns with computational efficiency. The combination of a stage-wise approach and dynamic attention head freezing preserves task-specific knowledge and prevents catastrophic forgetting, enabling the model to excel in both binary and multi-label classification tasks. Our innovative methodology supports effective knowledge transfer, ensures modular learning, and reduces task interference, particularly addressing the overlapping features of comorbid mental disorders. By leveraging continuous and transfer learning, the model adapts to evolving data without losing previously acquired knowledge, while guided training aligns attention heads with disorder-specific features, enhancing interpretability. These strategies result in a robust, scalable, and transparent framework, advancing the application of neural networks in mental health research and ensuring its suitability for complex clinical scenarios.

Appendix A. Supplementary Figures

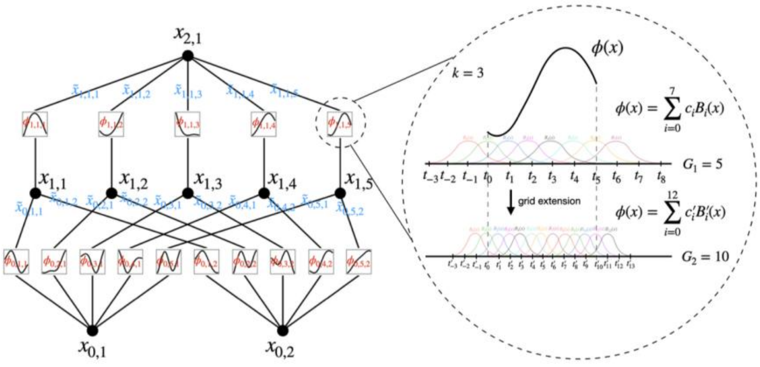

Figure A1. Notations of activations that flow through the network.

Right: an activation function is parameterized as a B-spline, which allows switching between coarse-grained and fine-grained grids. *KAN: Kolmogorov–Arnold Networks*

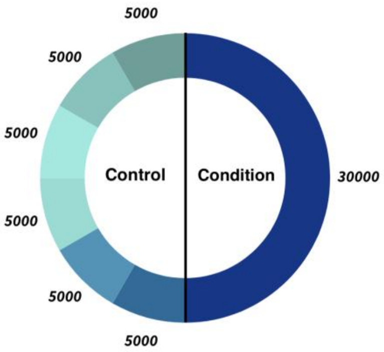

Figure A2. Barplot of case and control data for each mental health condition

Appendix B. Supplementary Tables

Table B1. Demonstration of the Depression Dataset

Text	**Label**
"I'm doing so so bad I am miserable. Today has been such a bad day. I'm so tired, disappointed with myself and sad. I am so overwhelmed I don't know how to control my thoughts."	**1**
"I'm worried about my ex-girlfriend? I was living with my girlfriend for 3 years. The past few months we were getting distant, I was dealing with lots of personal things so I began neglecting her a lot. ..didn't feel right leaving the door unlocked. It's been hours and no reply and she hasn't been on Facebook. I'm kinda worried about her."	0

This table provides an example of text data from the depression dataset. Each row includes a Reddit post and its corresponding label, where 1 indicates a mental health condition (depression), and 0 indicates a general post.

Table B2. Overlapping Conditions in Multilabel Dataset (53,044 Data)

Total Overlapping Cases	3,935
Percentage of Overlapping Cases	7.42%

Table B3. Demonstration of GPT Annotated Multilabel Dataset

Overlapping Conditions	**Number of Cases**
Depression + Anxiety	1,913
Depression + Bipolar	363
Depression + PTSD	301
Depression + Anxiety + PTSD	291
Depression + Anxiety + Bipolar	201

This table provides examples from the GPT-annotated multilabel dataset. Each entry contains a post and its corresponding mental health labels, where 1 indicates the presence of a condition, and 0 indicates its absence. This dataset facilitates multi-label classification tasks, enabling models to detect co-occurring mental health conditions.

Text	**Depression**	**Anxiety**	**OCD**	**Bipolar**	**PTSD**
"I feel so hopeless and like a failure. I do not know how to put this eloquently so I am just going to say it. This past year I applied to medical school (several different ones) and I got one interview which turned into a waitlist which I am still on. I feel like literally such a failure and I am burnt out and exhausted because I feel like I have worked so hard for four years just for my dreams that I have had for so long to be crushed... I just feel so sad and hopeless and depressed I do not know what to do."	**1**	0	0	0	**1**
"Never Going Home. So I have this thing, I don't go home all day. Once I wake up and get ready for the day I leave and go to my mom's. If she has something to do I'll go visit my aunt. If she has something to do I'll go visit a friend, so on and so forth. It is an avoidant tactic so I don't have to take care of things (cleaning, responding to mail, etc.)... As I'm typing this it seems like a healthy coping mechanism, but it drives me absolutely mad that I can't just take care of my shit."	0	0	0	**1**	0
"It gives you insomnia, which in turn makes your depression worse during the day, which messes up your sleep even more, which gives you anxiety. It just feels impossible to function. What is even worse is when I cannot sleep, my symptoms of my OCD and anxiety are just 50x worse. It feels impossible to escape depression."	**1**	**1**	**1**	0	0

Table B4. User-level predictions

Text Input	True Labels	Predicted Labels
"On paper I have a great life. Beautiful, smart fiancée. Rent an amazing apartment in the downtown of a great city. Have a wonderful dog. A well paying job. Lots of friends. Great family. But I'm always fucking miserable. Mostly during the work week. I don't know what it is. It's like Friday-Sunday I'm feeling great, hang out with friends, do fun stuff with my SO, party, watch sports, whatever. As soon as Monday hits, pretty much every little thing irritates me beyond belief. Stupid things. Basically if my routine or my 'expectations' of my routine get disrupted, I just shut down and I'm annoyed and silent all day. I catch myself in this mood a lot. And then I think about how stupid it is that I'm upset for no reason other than maybe I had to eat lunch 30 minutes later, or I had to walk the dog in the afternoon when I didn't plan on it, or I had to give up our office for my SO to work in for an hour or two. Things that are inconsequential. And then I get mad at myself for being so upset about nothing to the point I can't pull myself out of it. I recognize I'm in these moods but I just can't pull myself out of it. I shouldn't even get moody at these things to begin with."	Depression [1,0,0,0,0]	Depression [1,0,0,0,0]
"Why do I find the Spring so depressing? A lot of people talk about the Winter blues, but for me Spring by far (at least in recent years), is painfully depressing. It's like the smell in the air, mixed with the temperature and longer/brighter daytime present this fake sense of happiness. It's as if when I'm outside, things seem “too happy” and that scent in the air is gut-wrenchingly nostalgic of a past-time that can never be felt or experienced again. Because I'm not capable of it and I'm too worn-down to, anyway. I actually remember when I was little (29 now) that I loved clear, sunny weather. I looked up the forecast almost obsessively ahead of time, banking on those days of where there are no clouds to block the sun. Not even partly cloudy. It’s perplexing to think I was once like this since nowadays, I despise sunny weather. I genuinely feel better and more comfortable when it is cloudy with rain. Especially the eccentric types of weather where it looks as if nighttime has arrived too early, but instead it's just a storm brewing. Not to mention, I am at my peak mindset and performance late at night. What the hell happened."	OCD [0,0,1,0,0]	OCD [0,0,1,0,0]
"I just want to be happy and to make my partner happy. I don't understand why I'm like this. I love my partner more than anything, yet I struggle to think clearly and communicate effectively. I'm terrified that they're planning against me or will become tired of my episodes and end our relationship, finding me unbearably difficult. I find it hard to reach out for help because I'm unsure how to express my feelings without making them feel accused or thinking I'm losing my sanity. I simply want to gain control over myself, my emotions, and my thoughts. All I desire is to feel happy, to stop causing misery for everyone around me, and to share in the joy that others seem to experience. I feel overwhelmed by fear, anger, and confusion."	Bipolar [0,0,0,1,0]	Bipolar [0,0,0,1,0]
"(31f) I hate my life I know it just comes with trauma that I have no idea how to compact, and I feel so behind. I live at home with my mom because of student loans from a degree I had to drop out of because her credit score wasn't good enough, and neither was mine. I now sit with 80k in student debt and only 20k would be gone if Biden finally wipes away student debt.. My mom right now has been more unstable than before. I get it, I'm overweight, I have mental health issues, need some sun and a better job, but it doesn't help when she berates and complains about it daily and comparing me to others. I barely eat as it is, and while she serves unhealthy food as well, she gets mad that I'm not eating healthy and moving like a fucking swan. I'm like 200 lbs full of anxiety, different kinds of odd combinations of grass and veggies in some green smoothies that tastes like eating someone's ass that hasn't showered for 3 years. Still gets mad that I eat unhealthy when she makes it and it's literally all we have. She gets mad that I don't spend time with her at all and prefer to hang out with my friends that are online. She tells me I look ugly and I should look better in clothes that look ugly on me as it is. Literally, she treats me just like my older brother did minus the sexual abuse I endured for 14 fucking years (which ended when I was 26 by leaving to art school and finally having a way to make it end by severing ties with him (well he did it with me) Being yelled at because I get upset isn't a way to help someone unpack trauma nor help them get motivated about doing better. It's gotten so bad I can't focus on anything very well. I don't even have privacy to go and study to be a data analyst in	Depression, Anxiety, PTSD [1,1,0,0,1]	Anxiety, PTSD [0,1,0,0,1]

Coursera because school is really expensive nowadays and i don't have the time to be able to go. I feel really stuck. And I know many people are gonna say it's procrastination and I get it might be, but it stems from an overflowing has never stopped. I can't afford therapy because that shit isn't covered, nor can I drive to one because I don't have a car nor do I have the money to pay for an Uber drive weekly along with whatever fee therapy comes with.bi also never have privacy so I can't do at home therapy. I have so little privacy my mom barges in and tries to talk to me even though I tell her I'm in a literal meeting. But if I try to set boundaries or do things myself I'm called an asshole... It's so much thrown at me I feel like I just freeze and just sit and do nothing because that's better than sitting with her and possibly be yelled and berated at for my weight for the umpth time even though she's heavy and diabetic herself. Yeah. My live sucks right now..."		
"Hey. Thanks for reading. I am 22M from Europe and I was diagnosed with a significant mood disorder, obsessive tendencies, and high stress last year. I have also spent some time in a mental health care facility and was prescribed medications to manage my condition. The root of my struggles was a series of traumatic embarrassments and mistakes, most of which I caused myself. I've never been great at making decisions, even though I've read countless self-improvement books and other resources. Unfortunately, I've had a hard time retaining and applying what I learned, often acting impulsively or emotionally. To make things even more difficult, I was turned down by someone I deeply admired because of my behavior, which came across as awkward. She also discovered some things about me that I find embarrassing. I've been too open and vulnerable with people who didn't deserve it, and that hurts deeply. I dropped out of high school because my biggest passion was pursuing a music career (rap). However, I've been considering finishing my last year of school and going to college. It's partly because I want a backup plan in case my music career doesn't take off financially, but I also believe it might help with my struggles. My low mood is the real obstacle—it's keeping me from enjoying the things I love, trapping me in the past, and making me feel intense unease about what others might think of me. I never had this issue before. Is it possible to find joy again and become the athletic, quick-witted, confident, and happy person I was before my life took this turn? I also want to dedicate more time to the gym and writing lyrics, but I lack the drive to start. Can I pull myself out of this dark and challenging place and achieve my dreams? I hate feeling like I'm wasting my life, but I also can't simply "turn off" these feelings. Peace. 22 years old and struggling. Can I regain my confidence and happiness and find success in life?"	OCD, PTSD [0,0,1,0,1]	OCD, PTSD [0,0,1,0,1]

REFERENCE

Chen, S., Zhang, Z., Wu, M., & Zhu, K. (2023). Detection of multiple mental disorders from social media with two-stream psychiatric experts. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 9071-9084).

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019). Class-Balanced Loss Based on Effective Number of Samples. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Han, S., Mao, R., & Cambria, E. (2022). Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings. *arXiv preprint arXiv:2209.07494*.

Kirkpatrick, James, et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13), 3521-3526.

Liu, Ziming, et al. (2024). Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*.

Ruder, S. (2017). *An Overview of Multi-Task Learning in Deep Neural Networks.* arXiv preprint arXiv:1706.05098.

Rusu, Andrei A., et al. (2016). "Progressive neural networks." arXiv preprint arXiv:1606.04671.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762 .

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1810.04805

Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, Marsh J, DeVylder J, Walter M, Berrouiguet S, Lemey C. Machine Learning and Natural Language Processing in Mental Health: Systematic Review. J Med Internet Res. 2021 May 4;23(5):e15708. doi: 10.2196/15708. PMID: 33944788; PMCID: PMC8132982.

Lin, T. (2017). Focal Loss for Dense Object Detection. arXiv preprint arXiv:1708.02002.

Montejo-Ráez, A., Molina-González, M. D., Jiménez-Zafra, S. M., García-Cumbreras, M. Á., & García-López, L. J. (2024). A survey on detecting mental disorders with natural language processing: Literature review, trends and challenges. *Computer Science Review, 53,* 100654. https://doi.org/10.1016/j.cosrev.2024.100654

David Losada, Fabio Crestani, and Javier Parapar. 2017. erisk 2017: Clef lab on early risk prediction on the internet: Experimental foundations.

M. Roca, M. Gili, M. Garcia-Garcia, J. Salva, M. Vives, J. Garcia Campayo, and A. Comas. 2009. Prevalence and comorbidity of common mental disorders in primary care. Journal of Affective Disorders

Zhang, M.-L., & Zhou, Z.-H. (2014). A review on multi-label learning algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(1), 1–25. https://doi.org/10.1109/TPAMI.2013.140

Data Source:

https://academictorrents.com/details/56aa49f9653ba545f48df2e33679f014d2829c10

https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health

Embedding: bertmini Source

https://huggingface.co/prajjwal1/bert-mini